University Unit: Faculty of Computing and Informatics		
Course: Information Systems		***Thematic Core:***
Subject: **DATA SCIENCE**		***Subject Code:*** ENEX50117
Professor: **ARNALDO RABELLO de AGUIAR VALLIM FILHO**	***DRT:*** 10.4403-0	***Semester:*** **6**
Credit hours: 4 h/a (Theory - 2h In-person lectures and 2h Distance Learning)	(X) Theoretical ( ) Practice	***Academic Semester:*** 2024/2º

Syllabus:

Analysis of the knowledge discovery in databases (KDD). Machine Learning (ML). Supervised Learning: Classification and Regression. Unsupervised Learning: Clustering and Association. Text Representation and Mining. Introduction to Visualization.

Objectives:

Conceptual Objectives	***Procedural Objectives and Skills***	***Attitudinal Objectives and Values***
• Understand in general terms the Data Science and Analysis process and its applications • Understand the fundamentals and paradigms of ML models • Know the main ML models (development, applications, advantages and disadvantages of each model) • Know the main Data Analysis and ML Development Environments	• Evaluate Data Science methods and ML algorithms in Problem Solving and Decision Making • Build and execute Data Science/Analysis models in practical cases through ML algorithms	• Gain awareness of the potential of Data Science and ML and their limitations • Integrate the acquired skills in problem-solving and decision-making through the development of ML applications

Program Content:

1. **Introduction to Data Science**
   - 1.1. Big Data vs. Data Science
   - 1.2. The Process of Knowledge Discovery in Databases
   - 1.3. Data Science Techniques and Phases
2. **Exploratory Data Analysis**
   - 2.1. Review of Descriptive Statistics: Analysis Tools
   - 2.2 Review of Descriptive Statistics: Graphical Data Visualization Tools
3. **Data Preparation**
   - 3.1. Data Types
   - 3.2. Data Cleaning and Processing
   - 3.3. Attribute Selection
4. **Supervised Learning I: Classification Algorithms**
   - 4.1. Concept of Supervised Learning
   - 4.2. Classification by Decision Trees
   - 4.3. Classification by K-Nearest Neighbors
   - 4.4. Classification by the Naive Bayes Algorithm
   - 4.5. Model Training Strategies
   - 4.6. Model Evaluation Metrics
5. **Unsupervised Learning I: Clustering Algorithms**
   - 5.1. Concept of Unsupervised Learning
   - 5.2. K-Means Clustering
   - 5.3. Hierarchical Clustering
6. **Unsupervised Learning II: Association Rules**
   - 6.1. Data Association - Concepts
   - 6.2. Apriori Algorithm
   - 6.3. FP-Growth Algorithm
7. **Artificial Neural Networks (ANN)**
   - 7.1. ANN Concepts
   - 7.2. Neuron Model
   - 7.3. ANN Architecture
   - 7.4. MLP Network
   - 7.5. SOM Network
   - 7.6. Time Series Forecasting Using ANN
8. **Supervised Learning II: Regression Analysis through ML**
   - 8.1. Classical Regression Models (review)
   - 8.2. Model Evaluation (review)
   - 8.3. Linear Regression through ML
   - 8.4. Logistic Regression

Methodology:

In-person lectures
Case discussions
Active practices for content consolidation (individual and group)
EaD support, including development of activities, applications and video classes

Evaluation Criteria:

- **INTERMEDIATE EVALUATIONS (IE)**
  - **1st Semester Average (IE1)**
    - . Individual IE 1: 70% of N1
    - . Activities 1: 30% of N1
  - **2nd Semester Average (IE2)**
    - . Individual IE 2: 70% of N2
    - . Activities 2: 30% of N2

- **Participation Score (PS)**

  In the case of a Participation Score, there are two alternatives:
  - . If there is an Integrated Test ==> PS varies between 0.0 and 0.5
  - . If there is no Integrated Test ==> There is no PS (PS = 0.0)

- **IE Average (IEA)**

  IEA = (IE1 + IE2)/2 + PS

- **FINAL EXAM (FE)**

  Individual exam at the end of the semester

- ***SUBSTITUTE TEST (SUB)***

The student has the right to take a SUB at the end of the semester, which covers all the content of the subject. This grade may replace an Individual IE or an Activity that the student has missed. In case of having missed an Individual IE and one or more Activities, the SUB grade replaces the Individual IE grade. In case of having missed more than one Activity, and no Individual IE, the SUB grade replaces the Activity with the highest Weight, if there is a difference in the Weights..

CRITÉRIO de APROVAÇÃO

If: *Attendance* ≥ *75%* e IEA ≥ 6.0 ==> APPROVED and exempt from FE

- **FINAL AVERAGE (FA) = IEA**

Otherwise: ==> Student takes FE, if Attendance ≥ 75%

- **FA = 0.5 IEA + 0.5FE**

If: FA ≥ 6,0 ==> APPROVED

In any other case: Student has FAILED

Basic Bibliography:

FACELI, K., LORENA, A. C. ; GAMA, J. ; CARVALHO, A. C. P. L. F. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. 1. ed., Editora LTC, 2011.

HAN, J.; KAMBER, M. ; PEI, J. (2012) **Data Mining, Concepts and Techniques**. 3rd Edition. Elsevier - Morgan Kauffman.

SILVA, A.S.; PERES, S.M.; BOSCARIOLI, C. I**ntrodução a Mineração de dados – Com aplicações em R.** 1ª. ed., Editora Elsevier, 2016.

Additional Bibliography:

DEVORE, J. L. Probabilidade e estatística para Engenharia e Ciências. São Paulo. CENGAGE Learning. 2011.

LANTZ, BRETT. **Machine Learning wiht R**. Packt Publishing Open Source. 2015. 2ª Ed. ISBN 978-1-78439-390-8. [novo]

MOORE, D. S.; NOTZ, W. I.; FLINGER, M. A. A Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC, 2017.

R Documentation. https://www.r-project.org/.

WITTEN I. H., EIBE F., MARK A. H. **Data Mining: Practical Machine Learning Tools and Techniques**, 3a. ed., Editora Morgan Kaufmann, 2011.

ZHAO, Yanchang. **R and Data Mining: Examples and Case Studies.** Disponível em: http://www.RDataMining.com.